

Inferring Disease and Gene Set Associations with Rank Coherence in Networks

TaeHyun Hwang¹, Wei Zhang¹, Maoqiang Xie², Rui Kuang^{1,*}

1 Department of Computer Science and Engineering, University of Minnesota Twin Cities, MN, USA

2 College of Software, Nankai University, Tianjin, China

* Correspondence: kuang@cs.umn.edu

Abstract

A computational challenge to validate the candidate disease genes identified in a high-throughput genomic study is to elucidate the associations between the set of candidate genes and disease phenotypes. The conventional gene set enrichment analysis often fails to reveal associations between disease phenotypes and the gene sets with a short list of poorly annotated genes, because the existing annotations of disease causative genes are incomplete. We propose a network-based computational approach called rcNet to discover the associations between gene sets and disease phenotypes. Assuming coherent associations between the genes ranked by their relevance to the query gene set, and the disease phenotypes ranked by their relevance to the hidden target disease phenotypes of the query gene set, we formulate a learning framework maximizing the rank coherence with respect to the known disease phenotype-gene associations. An efficient algorithm coupling ridge regression with label propagation, and two variants are introduced to find the optimal solution of the framework. We evaluated the rcNet algorithms and existing baseline methods with both leave-one-out cross-validation and a task of predicting recently discovered disease-gene associations in OMIM. The experiments demonstrated that the rcNet algorithms achieved the best overall rankings compared to the baselines. To further validate the reproducibility of the performance, we applied the algorithms to identify the target diseases of novel candidate disease genes obtained from recent studies of GWAS, DNA copy number variation analysis, and gene expression profiling. The algorithms ranked the target disease of the candidate genes at the top of the rank list in many cases across all the three case studies. The rcNet algorithms are available as a webtool for disease and gene set association analysis at http://compbio.cs.umn.edu/dgsa_rcNet.

Author Summary

Introduction

Determination of the molecular cause of diseases is a major focus in genomics research since early 1960s [1]. Recently, powered by the advanced high-throughput genomic technologies, numerous large-scale genome-wide disease studies such as genome-wide association studies [2,3], DNA copy number detections [4], and gene expression profiling [5], were conducted towards this goal. Typically, the objective of a study is to perform a high-throughput scanning for a list of genes that are involved with the disease under study, and then a standard follow-up enrichment analysis or its variants and extensions is applied to analyze the gene set, based on the statistical significance of the overlap between the genes and gene functional annotations or associations with disease phenotypes. Examples of the well-known tools are DAVID [6], GSEA [7], GOToolBox [8] and many others. However, in many cases, since the existing annotations of disease causative genes is far from complete [1], and a gene set might only contain a short list of poorly annotated genes, enrichment-based approaches often fail to reveal the associations between gene sets and disease phenotypes.

The availability of large phenotypic and molecular networks provides a new opportunity to study the association between diseases and the gene sets identified from the high-throughput genomic studies. The

human disease phenotype network [9] provides information on phenotype similarities computed by text mining of the full text and clinical synopsis of the disease phenotypes in OMIM [1]. Large molecular networks such as the human protein-protein interaction network [10] or functional linkage network [11] provide functional relations among genes or proteins. Based on the observation that genes associated with the same or related diseases tend to interact with each other in the gene network, many network-based approaches are proposed to utilize the disease modules and gene modules in the networks to prioritize disease genes, a task of ranking genes for studying genetic diseases [10–16].

In this paper, we propose a general network-based approach to infer associations between disease phenotypes and gene sets, utilizing the disease phenotype network and the gene network. We formulate the problem as a gene set query problem. By querying the networks with a given gene set, a user expects to retrieve a list of disease phenotypes with the highest predicted association with the gene set. Based on the assumption that the genes ranked by their relevance to the query gene set will have coherent associations with the disease phenotypes ranked by their relevance to the hidden target disease phenotypes, we formulate a simple learning framework maximizing Rank Coherence in Networks (rcNet) with respect to the known disease phenotype-gene associations in OMIM. Fig. 1 illustrates the general idea of Rank Coherence in Networks. We first measure the global relevance between the query gene set and all the genes with graph Laplacian scores (Fig. 1A&B). The Laplacian scores can be considered as the result of using the query gene set as the seed to perform random walk with restart (or label propagation) in the gene network [17]. The global relevance between a target disease phenotype and all disease phenotypes can be similarly computed as the Laplacian scores with random walk on the disease phenotype network (Fig. 1D). Our assumption is that, between the rankings given by the query gene set and the target disease phenotype, the top-ranked genes and the top-ranked phenotypes should be highly connected by known associations, quantified by Rank Coherence in Networks (Fig. 1C). In a real problem, the target disease phenotypes are unknown. The rcNet algorithms are designed to search for the phenotype(s) with the best rcNet score against the query gene set. We propose two strategies. The first approach relaxes the combinatorial problem as ridge regression to find a closed-form solution for selecting the target disease phenotype. The second approach in two variants enumerates all possible phenotype configurations to find the best match of the query gene set.

The rcNet algorithms are different from the gene set enrichment analysis with statistical methods such as Hypergeometric statistics, permutation test or non-parametric McNemar’s test [6–8] because the rcNet algorithms use the topological information in the disease phenotype network and the gene network to analyze the association between a gene set and all phenotypes simultaneously. The simultaneous analysis of all phenotypes provides a global dependence, and thus richer and more reliable information for computing the association scores are used to rank the phenotypes. The rcNet algorithms share more algorithmic similarity with the disease gene prioritization methods, which were proposed for a different purpose. CIPHER [10] scores each gene against a disease phenotype based on the correlation between their relevances with all the phenotypes, where the relevance between the gene and a phenotype is calculated based on the distance between the gene and the genes associated with the phenotype. The methods proposed by [13], [15] and [11] applied random walk (label propagation) or simpler neighborhood weighting to exploit the gene networks for ranking genes for a disease phenotype, based on the seed genes mapped from the disease phenotype. One limitation is that the phenotype network and the sparse known associations are not fully utilized in the global analysis. The label propagation algorithms proposed by [14] and [16] explore a heterogeneous network combining the gene network, the phenotype network and the associations to explore gene modules, phenotype modules and the phenotype-gene association biclusters. Since the two methods make full use of the information in the networks, it is difficult to interpret the results and to tune the best parameters for combining the information.

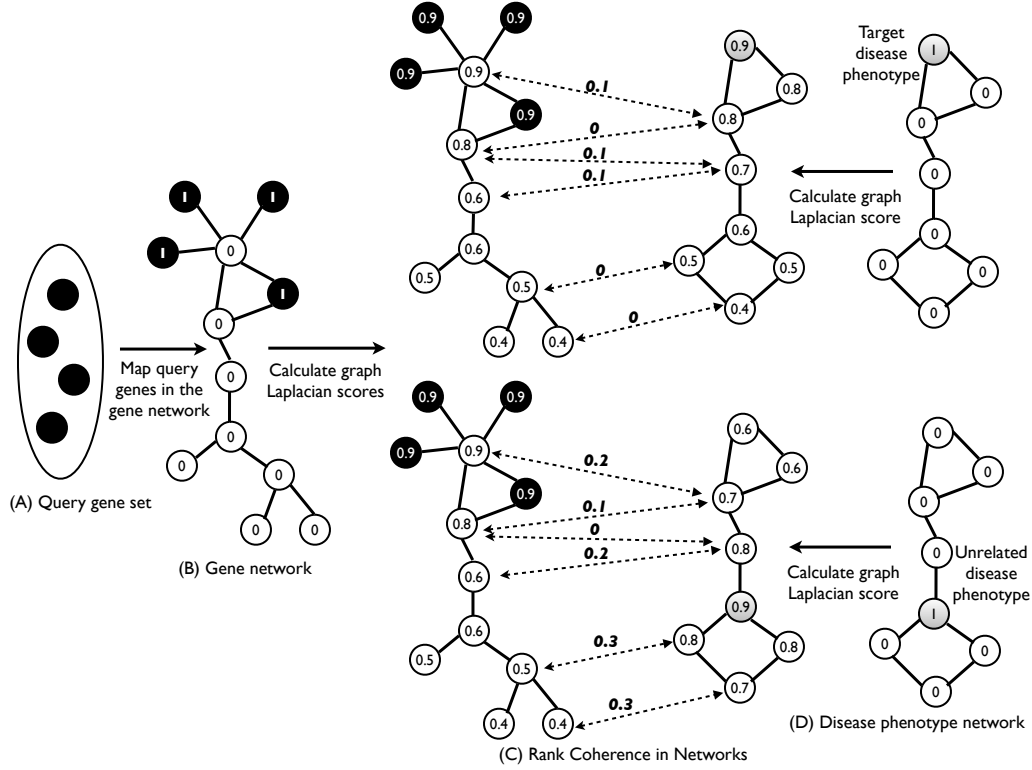


Figure 1. Illustration of Rank Coherence in Networks. A query gene set of four genes is given in (A). The four genes are mapped in the gene network and the corresponding nodes are marked with 1 in (B). The graph Laplacian scores are then computed to quantify the relevance between each gene (including the query genes) and the query gene set. Similarly, if a disease phenotype of the gene set is selected and marked with 1, the graph Laplacian scores can be derived to quantify the relevance between each disease phenotype and the selected phenotype in (D). Based on the coherence assumption, the top-ranked genes and the top-ranked phenotypes should be highly connected with each other if the phenotype is the target of the query gene set, otherwise the connectivity will be close to random. As showed in (C), the edges connecting associated genes and phenotypes are labeled by the discrepancy between their ranking scores. Clearly, the phenotype ranking given by target phenotype query is more coherent (the upper case) than the ranking given by an unrelated phenotype (the bottom case). The connectivity is measured by Rank Coherence in the Networks (rcNet). In general, since the target disease phenotypes are not known, the rcNet algorithms search for the phenotype with the best rcNet score against the query gene set.

```

dgsa.rcNet(g,  $\bar{\mathbf{G}}$ ,  $\bar{\mathbf{P}}$ , A,  $\alpha$ ,  $\beta$ )

1 p = 0

2  $\tilde{\mathbf{g}} = (\mathbf{1} - \alpha)(\mathbf{I} - \alpha\bar{\mathbf{G}})^{-1}\mathbf{g}$  (equation (3)).

3  $\bar{\mathbf{A}} = (\mathbf{1} - \beta)\mathbf{A}(\mathbf{I} - \beta\bar{\mathbf{P}})^{-1}$ 

4  $\mathbf{p}^* = (\bar{\mathbf{A}}^T\bar{\mathbf{A}} + \kappa\mathbf{I})^{-1}\bar{\mathbf{A}}^T\tilde{\mathbf{g}}$ 

5  $\mathbf{p}(\mathbf{p}^* > \mathbf{a}) = \mathbf{1}$  (target selection with threshold a)

6 return (p)

```

Figure 2. rcNet Algorithm - Rank Coherence in Networks.

Methods

Problem Definition

We formulate a graph query problem for disease phenotype and gene set association discovery: given a heterogenous network consisting of the gene network, the phenotype network and the association network, we query the network with a gene set to retrieve a phenotype (or several) predicted to have association with the query gene set. We define $\mathbf{G}_{(\mathbf{n} \times \mathbf{n})}$, $\mathbf{P}_{(\mathbf{m} \times \mathbf{m})}$, and $\mathbf{A}_{(\mathbf{n} \times \mathbf{m})}$ as the adjacency matrix of the gene network, the disease network, and the disease-gene association network, respectively, where \mathbf{n} is the number of genes and \mathbf{m} is the number of disease phenotypes in the networks. The query gene set is represented by a binary vector $\mathbf{g} = [\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_n]^T$ denoting the gene membership against the gene set, i.e. each $\mathbf{g}_i = \mathbf{1}$ if gene i is in the query gene set, otherwise $\mathbf{0}$. Similarly, the list of target phenotype(s) is given by another binary vector $\mathbf{p} = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_m]^T$ and phenotype \mathbf{j} is a target phenotype if $\mathbf{p}_j = \mathbf{1}$. Our objective is to find the \mathbf{p} that gives the best rank coherence with the query gene set \mathbf{g} .

Computing Graph Laplacian Scores

To fully utilize network topological information, we compute the global relevance score between the query gene set \mathbf{g} and all the genes based on the graph Laplacian of the gene network $\mathbf{G}_{(\mathbf{n} \times \mathbf{n})}$. We first normalize \mathbf{G} as $\bar{\mathbf{G}} = \mathbf{D}_G^{\frac{1}{2}} \mathbf{G} \mathbf{D}_G^{\frac{1}{2}}$, where \mathbf{D}_G is a diagonal matrix with diagonal elements $\mathbf{D}_{G\mathbf{i},\mathbf{i}} = \sum_j \mathbf{G}_{\mathbf{i},j}$. A vector $\tilde{\mathbf{g}}$ of graph Laplacian scores is derived from the following optimization problem [18],

$$\min_{\tilde{\mathbf{g}}} \sum_{\mathbf{i}, \mathbf{j}} \bar{\mathbf{G}}_{\mathbf{i}, \mathbf{j}} (\tilde{\mathbf{g}}_{\mathbf{i}} - \tilde{\mathbf{g}}_{\mathbf{j}})^2 + \frac{1 - \alpha}{\alpha} \sum_{\mathbf{i}} (\tilde{\mathbf{g}}_{\mathbf{i}} - \mathbf{g}_{\mathbf{i}})^2 \quad (1)$$

In equation (1), the first term is a smoothness penalty, which forces connected genes to receive similar scores, and the second term ensures the consistency with the query gene set. The Laplacian scores combine the neighboring information in the network with the consistency with the query gene set to provide a global relevance measure between each gene and the query gene set. Parameter $\alpha \in (0, 1)$ balances the contributions from the two penalties. The closed-form solution of equation (1) is

$$\tilde{\mathbf{g}} = (\mathbf{1} - \alpha)(\mathbf{I} - \alpha\bar{\mathbf{G}})^{-1}\mathbf{g}. \quad (2)$$

Empirically, to avoid computing the inverse of $(\mathbf{I} - \alpha\bar{\mathbf{G}})$, an iterative algorithm can efficiently compute the closed-form solution with the following update rule at each time step t ,

$$\tilde{\mathbf{g}}^t = (\mathbf{1} - \alpha)\mathbf{g} + \alpha\bar{\mathbf{G}}\tilde{\mathbf{g}}^{t-1}, \quad (3)$$

Similarly, graph Laplacian scores can be derived to measure the relevance between the phenotypes and the target phenotypes \mathbf{p} with optimization of

$$\min_{\tilde{\mathbf{p}}} \sum_{i,j} \bar{\mathbf{P}}_{i,j} (\tilde{\mathbf{p}}_i - \tilde{\mathbf{p}}_j)^2 + \frac{1-\beta}{\beta} \sum_i (\tilde{\mathbf{p}}_i - \mathbf{p}_i)^2, \quad (4)$$

with the closed-form solution

$$\tilde{\mathbf{p}} = (\mathbf{1} - \beta)(\mathbf{I} - \beta\bar{\mathbf{P}})^{-1}\mathbf{p}, \quad (5)$$

where $\bar{\mathbf{P}}$ is the normalized \mathbf{P} and $\beta \in (0, 1)$ is the balancing parameter. Computing the laplacian scores is equivalent to a weighted summation of performing random walk on the graph with all the steps to infinite. Thus, the laplacian scores exploit modular information in a network to capture long range interactions between the nodes in a graph. Note that one can use other scoring functions such as counting the direct neighbors of the query gene set, or measuring the shortest path from the query gene set to other genes as suggested in [10]. However, empirically, the direct-neighbor function tends to generate very sparse information, and the shortest-path function does not fully explore the neighborhood information.

Rank Coherence in Networks

Rank Coherence in Networks (rcNet) measures whether the query gene set \mathbf{g} and a phenotype set \mathbf{p} show coherent associations with the known disease-gene associations. Specifically, given the graph Laplacian scores $\tilde{\mathbf{g}}$, which rank the genes by their relevance to the query gene set \mathbf{g} , and the graph Laplacian scores $\tilde{\mathbf{p}}$, which rank the disease phenotypes by their relevance to the hidden target phenotypes \mathbf{p} , Rank Coherence in Networks $\text{rcNet}(\tilde{\mathbf{g}}, \tilde{\mathbf{p}}, \mathbf{A})$ measures whether the associations given by \mathbf{A} are connecting genes and phenotypes with similar scores in $\tilde{\mathbf{g}}$ and $\tilde{\mathbf{p}}$. We propose two different approaches to define Rank Coherence in Networks. The first approach adopts a ridge regression model coupled with label propagations to compute a closed-form solution of \mathbf{p} , relaxed to real numbers. The second approach uses simpler measures and enumerate all possible \mathbf{p} to find the best fitting for \mathbf{g} .

A Ridge Regression Model

Under the assumption that the Laplacian score of a phenotype can be reconstructed by the linear combination of the Laplacian scores of its gene neighbors in \mathbf{A} , we can formulate the following least-square cost function,

$$\Omega = \|\mathbf{A}\tilde{\mathbf{p}} - \tilde{\mathbf{g}}\|^2. \quad (6)$$

Eventually, we are interested in deriving \mathbf{p} . After replacing $\tilde{\mathbf{g}}$ with equation (2) and $\tilde{\mathbf{p}}$ with equation (5), we have the following regularization framework,

$$\Omega(\mathbf{p}) = \|(\mathbf{1} - \beta)\mathbf{A}(\mathbf{I} - \beta\bar{\mathbf{P}})^{-1}\mathbf{p} - (\mathbf{1} - \alpha)(\mathbf{I} - \alpha\bar{\mathbf{G}})^{-1}\mathbf{g}\|^2 + \kappa\|\mathbf{p}\|^2, \quad (7)$$

where $\|\mathbf{p}\|^2$ is a 2-norm regularizer and κ is a small constant. Equation (7) takes the standard form of ridge regression, and thus the closed-form solution \mathbf{p}^* can be derived by

$$\mathbf{p}^* = (\mathbf{1} - \alpha)(\bar{\mathbf{A}}^T\bar{\mathbf{A}} + \kappa\mathbf{I})^{-1}\bar{\mathbf{A}}^T(\mathbf{I} - \alpha\bar{\mathbf{G}})^{-1}\mathbf{g}. \quad (8)$$

where $\bar{\mathbf{A}} = (\mathbf{1} - \beta)\mathbf{A}(\mathbf{I} - \beta\bar{\mathbf{P}})^{-1}$. Note that the solution \mathbf{p}^* is a real vector, which can be seen as an approximation of the binary vector \mathbf{p} . A simple post-processing is to select one or a few phenotypes that are assigned with significantly larger scores as the phenotypes associated with the gene set. The full algorithm to solve the ridge regression model is given in Fig. 2. The steps at line 2, 3 and 4 require cubic matrix inversion algorithms. Thus, the time complexity of rcNet algorithm is $\mathcal{O}(\mathbf{m}^3 + \mathbf{n}^3)$.

```

dgsa.rcNet_enu(g,  $\bar{\mathbf{G}}$ ,  $\bar{\mathbf{P}}$ , A,  $\alpha$ ,  $\beta$ )

1  $\tilde{\mathbf{g}} = (\mathbf{I} - \alpha)(\mathbf{I} - \alpha\bar{\mathbf{G}})^{-1}\mathbf{g}$ 
2  $\mathbf{p} = \mathbf{0}, \mathbf{s} = \mathbf{0}$ 
3 for  $i = 1$  to  $n$ 
4    $\mathbf{p}_i = 1$ 
5    $\tilde{\mathbf{p}} = (\mathbf{I} - \beta)(\mathbf{I} - \beta\bar{\mathbf{P}})^{-1}\mathbf{p}$ .
6    $\mathbf{s}_i = \text{corr}(\mathbf{A}\tilde{\mathbf{p}}, \tilde{\mathbf{g}})$     or
    $-\sum_{i,j} \mathbf{A}_{i,j}(\tilde{\mathbf{p}}_i - \tilde{\mathbf{g}}_j)^2$ 
7    $\mathbf{p}_i = \mathbf{0}$ 
8    $j = \text{argmax}_i \mathbf{s}_i$ 
9    $\mathbf{p}_j = 1$ 
10 return ( $\mathbf{p}$ )

```

Figure 3. $\text{rcNet}_{\text{corr}}$ and $\text{rcNet}_{\text{lap}}$ Algorithms - Rank Coherence in Networks by Enumeration.

Enumeration Methods

The ridge regression model provides an approximation solution, but if we are only interested in retrieving the most relevant disease phenotype. We can simply go through each phenotype and compute a score against the query gene set \mathbf{g} for each case. Finally, the phenotype with the largest score is chosen as the target phenotype. We propose two functions to measure rcNet for this approach,

$$\text{rcNet}_{\text{corr}}(\tilde{\mathbf{g}}, \tilde{\mathbf{p}}, \mathbf{A}) = \text{corr}(\mathbf{A}\tilde{\mathbf{p}}, \tilde{\mathbf{g}}), \quad (9)$$

$$\text{rcNet}_{\text{lap}}(\tilde{\mathbf{g}}, \tilde{\mathbf{p}}, \mathbf{A}) = -\sum_{i,j} \mathbf{A}_{i,j}(\tilde{\mathbf{p}}_i - \tilde{\mathbf{g}}_j)^2. \quad (10)$$

Function $\text{rcNet}_{\text{corr}}$ simply uses the Pearson correlation coefficient to check the consistency between $\mathbf{A}\tilde{\mathbf{p}}$ and $\tilde{\mathbf{g}}$, similar to the concordance score used by CIPHER [10]. Function $\text{rcNet}_{\text{lap}}$ checks if the neighboring genes and phenotypes in the association network are assigned similar scores, and the smaller the disagreement, the higher the relevance. This enumeration strategy is similar to CIPHER [10]. The advantages are the conceptual simplicity and the optimality of the exact solution. The disadvantages are the computational cost incurred by the repeated calculation of the association score for each possible combination of the individual phenotypes, and the inflexibility to extend to more general problem of finding multiple target phenotypes. The full algorithm to solve the two enumeration models is given in Fig. 3. Inside the for-loop between line 3 and 7, the rcNet score is computed for each configuration of \mathbf{p} . The overall time complexity of this algorithm is also $\mathcal{O}(\mathbf{m}^3 + \mathbf{n}^3)$ if $(\mathbf{I} - \beta)(\mathbf{I} - \beta\bar{\mathbf{P}})^{-1}$ is precomputed. Note that this is the computational cost by which we only want to retrieve one phenotype. If we want to explore all possible configurations of \mathbf{p} , the total cost is exponential in \mathbf{m} .

Results

The rcNet algorithms are first compared to other methods in experiments of leave-one-out cross-validation and a task of predicting recently discovered disease-gene associations with OMIM data. The rcNet

algorithms are then applied to validate findings in datasets from GWAS, DNA copy number analysis, and microarray gene expression profiling.

Preparing Networks

The disease phenotype network is an undirected graph with 5080 vertices representing OMIM disease phenotypes, and edges with weights in $[0, 1]$. The edge weights measure the similarity between two phenotypes by their overlap in the text and the clinical synopsis in OMIM records, calculated by text mining [9].

The disease-gene associations are represented by an undirected bipartite graph with edges connecting phenotype nodes with their causative gene nodes. Two versions (May-2007 Version and May-2010 Version) of OMIM associations were used in the experiments [1]. The May-2007 Version contains 1393 associations between 1126 disease phenotypes and 916 genes, and the May-2010 Version contains 2469 associations connecting 1786 disease phenotypes and 1636 genes. The May-2007 version was used in the validation experiments on the OMIM data and the GWAS datasets, and the May-2010 version was used in the experiments on the DNA copy number and gene expression datasets.

Two gene networks were used in the experiments. The first one was derived from the human protein-protein interaction (PPI) network obtained from HPRD [19]. The PPI network contains 34,364 binary undirected interactions between 8919 genes. This network was used in the experiments on the OMIM data. A larger human functional linkage network [20] was used in the experiments on the GWAS, DNA copy number and gene expression datasets. This network contains 24,433 genes and around 60 million weighted edges. To reduce the computational complexity, we applied a cutoff 0.6 on the edge weights to generate a sparser network with around 7 million weighted edges.

Comparison with Other Methods and Evaluations

The rcNet algorithms were compared with CIPHER [10] and Random Walk with Restart (label propagation) methods [13–16], since those methods reported the best performance for disease gene prioritization. We adopted CIPHER with direct neighbor (C-DN) or shortest path (C-SP) for disease phenotype and gene set association analysis by averaging the correlations across the genes in the query gene set. The Random Walk algorithm described in [16] (RWR) was chosen as the label propagation method for comparison because it is straightforward to use the model for disease phenotype and gene set association analysis. The two hyper-parameters α and β for rcNet were chosen from $\{0.1, 0.5, 0.9\}$, and a small number $\kappa = 10^{-5}$ was chosen for ridge regression in all experiments. The three balancing parameters for RWR were also chosen from $\{0.1, 0.5, 0.9\}$. For all the methods, the results produced by the best parameters in the leave-one-out cross-validation were reported.

In all the experiments, a query gene set was used to rank all the 5080 disease phenotypes. The higher the target phenotype in the ranking, the better the performance. We measured the performance of a method with receiver operating characteristic (ROC) score, also called area under curve (AUC). Since we are most interested in whether the target phenotype is near the top, we report the area under the ROC curve up to the first 50 and 100 false positives. Another important evaluation is how well a method selects highly coherent top-ranked genes and top-ranked phenotypes since high coherence implies a good utilization of known associations in the model. Specifically, the top genes and phenotypes ranked by the query gene set and the target disease phenotype are assigned largest scores in the cost functions, and connections cancels out the large scores to give smaller penalty. To quantify the connectivity, the top- \mathbf{r} disease genes and the top- \mathbf{l} disease phenotypes with known OMIM disease-gene associations are selected to measure *fold enrichment*, which is calculated as $\frac{\mathbf{k}}{(\mathbf{r} * \mathbf{l}) * \mathbf{e}}$, where \mathbf{k} is the number of observed OMIM associations between the \mathbf{r} genes and the \mathbf{l} disease phenotypes, and \mathbf{e} is the probability of observing a random association between a gene node and phenotype node, estimated from the density of the OMIM disease phenotype-gene associations. Higher fold enrichment indicates higher coherence between

Table 1. Performance comparison in leave-one-out cross-validation and new association prediction with OMIM data. The tables report the average ROC_{50} and ROC_{100} across all the query cases for each method.

(A) Leave-one-out cross-validation						
Methods	rcNet	rcNet _{corr}	rcNet _{lap}	RWR	CDN	CSP
ROC_{50}	0.160	0.195	0.198	0.140	0.139	0.154
ROC_{100}	0.206	0.254	0.257	0.193	0.197	0.195

(B) Prediction of novel disease phenotype-gene associations						
Methods	rcNet	rcNet _{corr}	rcNet _{lap}	RWR	CDN	CSP
ROC_{50}	0.117	0.134	0.136	0.110	0.077	0.062
ROC_{100}	0.151	0.177	0.178	0.148	0.103	0.096

top-ranked genes and disease phenotypes, i.e. highly connected with OMIM disease phenotype-gene associations.

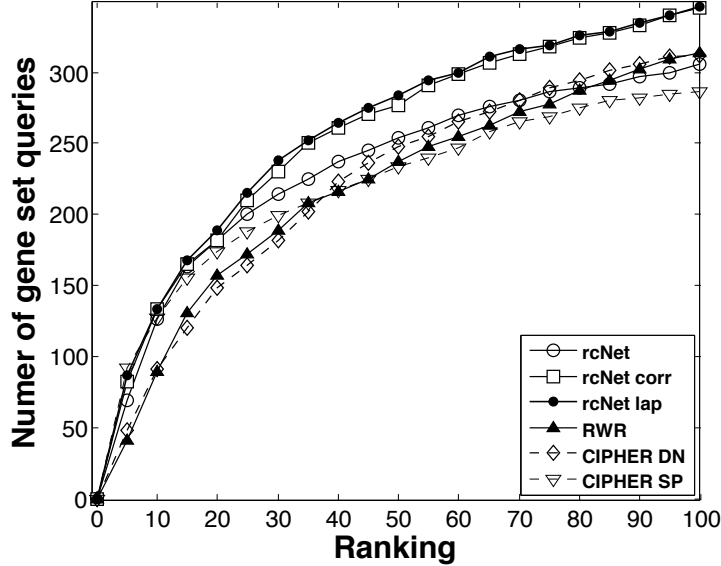


Figure 4. Ranking comparison in leave-one-out cross-validation. This figure reports the number of query cases, on which a method ranked the target disease phenotype among the top $k \in [1, 100]$ phenotypes.

Leave-one-out Cross-validation in OMIM

For each disease phenotype, the genes associated with the phenotype in OMIM were used as the query gene set to retrieve the disease phenotype. Note that the associations between the query gene set and all disease phenotypes including the target disease phenotype were removed in the experiment for leave-one-out cross-validation. In the experiments with RWR, as suggested by [16], the disease phenotype network was pruned by taking the 5 nearest neighbors of each node to reduce the computational complexity in leave-one-out cross-validation. Table 1(A) reports the average ROC_{50} and ROC_{100} scores across all the query cases in the leave-one-out cross-validation. Overall, the rcNet algorithms outperformed the other methods. Specifically, rcNet_{corr} and rcNet_{lap} achieved the best results with about 5% and 6% better ranking compared with the best of the others. rcNet performed slightly better than RWR, while CIPHER DN and CIPHER SP achieved lower scores. Fig. 4 shows a global comparison of the ranking by plotting the number of query cases with the target disease phenotype ranked above a certain rank. Clearly, the rcNet algorithms achieved better rankings at any ranking threshold in the experiments. For example, rcNet_{corr} and rcNet_{lap} ranked around 290 query cases above rank 50, while RWR and CIPHERs

ranked around 230 query cases above the rank.

We further analyzed how the rank coherence between the rankings by the query gene set and the disease phenotypes could affect the performance of rcNet algorithms and CIPHER. Based on the coherence assumption, the top ranked genes and the top ranked phenotypes by the query gene set and the target disease should be highly connected with each other by a good method. Fig. 5 compares the number of queries that achieved a significant fold enrichment for the target disease phenotype compared with the unrelated phenotypes. rcNet consistently identified more cases with significant fold enrichment against CIPHER SP at all the z -score thresholds. This observation suggests that label propagation is a better measure than shortest path to distinguish a target phenotype from unrelated ones because the information of gene neighborhoods are better utilized. Interestingly, CIPHER DN detected much less associations with high significance, but more cases with very high significance. For some query gene sets, which include genes with many disease genes as direct neighbors in dense disease gene modules, the direct gene neighbors tend to have more dense associations with the related phenotypes. However, only less than one-third of the query gene sets are the easy cases. CIPHER DN failed to find a significant fold enrichment for the other two-thirds. Thus, CIPHER DN is not performing well in general.

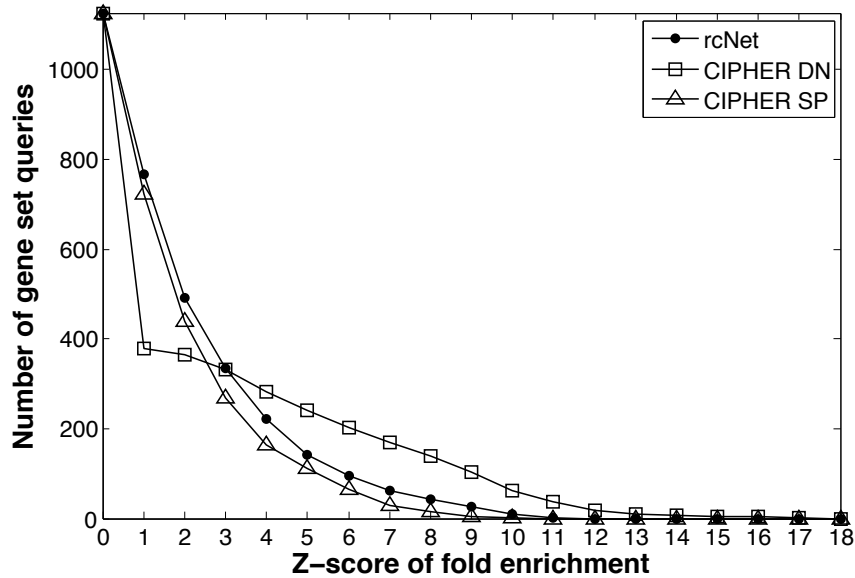


Figure 5. Fold enrichment significance. For each query, the association fold enrichment between the top-20 genes ranked by the query gene set and the top-20 disease phenotypes ranked by each disease phenotype is calculated. A z -score of fold enrichment was then computed for the target disease phenotype based on the scores of the unrelated disease phenotypes. This figure plots the number of query cases with a z -score above varying thresholds. A z -score at 1.96 corresponds to a p -value=0.05, assuming a normal distribution.

Predicting new OMIM Associations

To further evaluate how well a method can predict new disease-gene associations based on known associations, a case study was designed to predict the target disease phenotype of the new disease genes added into OMIM between May, 2007 and May, 2010. There are 387 new disease phenotypes were annotated in OMIM since May, 2007, excluding 11 new disease phenotypes whose disease genes have no interaction in the gene network. In this experiment, the task is to predict the target disease phenotype of the newly annotated disease genes, i.e. to query a set of new disease genes of a disease phenotype to retrieve the phenotype based on the disease-gene associations in May-2007 Version. Table 1(B) reports the average ROC_{50} and the ROC_{100} scores. $rcNet_{corr}$ and $rcNet_{lap}$ performed the best, followed by rcNet and RWR,

and CIPHER DN and CIPHER SP did not produce comparable results with the other methods. A possible reason for the worse performance of CIPHER is that the new cases are relatively under studied compared with the other disease phenotypes, and the global information in all the networks are necessary for an accurate inference of the associations. The results further supports the better performance of the rcNet algorithms compared with the baselines.

Table 2. Ranking the target disease phenotype of the disease susceptibility genes identified from GWAS. The disease categories in the first column are based on the definition in [21]. In the third column, the PubMed IDs marked with ‘*’ denote multiple GWASs for a disease/trait. Refer to supplementary Table for the results of the full list of the GWAS cases.

Category	Disease/Trait	OMIM Index	Gene Set Size	Rank by rcNet	Rank by rcNet _{corr}	Rank by rcNet _{lap}
Cancer	Prostate cancer	176807	15	2 (0.03%)	2 (0.03%)	2 (0.03%)
	Breast cancer	113705	26	7 (0.1%)	51 (1%)	43 (0.8%)
	Basal cell carcinoma (cutaneous)	605462	5	7 (0.1%)	189 (3.7%)	228 (4.5%)
	Basal cell carcinoma (cutaneous)	604451	5	90 (2%)	202 (4%)	256 (5%)
	Urinary bladder cancer	109800	1	14 (0.2%)	48 (0.9%)	60 (1.1%)
	Acute lymphoblastic leukemia (childhood)	159555	3	19 (0.04%)	51 (1.0%)	45 (0.8%)
	Lung cancer	211980	12	22 (0.4%)	587 (12%)	1610 (32%)
	Lung adenocarcinoma	211980	6	52 (1%)	838 (16%)	1815 (36%)
	Chronic lymphocytic leukemia	151430	14	57 (1%)	318 (6.3%)	306 (6%)
	Neuroblastoma (high-risk)	600613	1	143 (3%)	110 (2%)	138 (3%)
Immunological	Systemic lupus erythematosus	152700	10	46 (0.9%)	178 (4%)	161 (3%)
	Leprosy	246300	4	78 (1.5%)	62 (1.2%)	64 (1.3%)
	Leprosy	607572	4	272 (5%)	54 (1%)	55 (1%)
Endocrine	Type 2 diabetes	125853	9	97 (2%)	718 (14%)	1912 (38%)
	Type 1 diabetes	222100	26	331 (7%)	690 (13%)	191 (3.8%)
Gastrointestinal	Crohns disease	266600	2	60 (1.2%)	1396 (27%)	3012 (59%)

Predicting Disease Phenotypes of Disease Susceptibility Genes from GWAS

The goal of GWAS is to discover disease susceptibility loci/genes that could be useful for assessing or predicting an individual’s risk of disease. However, it is often challenging to assess how a set of novel disease susceptibility genes potentially influence susceptibility in disease, especially when the set of genes have no or little previously known disease implications, or function and pathway annotations. In this case study, we collected new disease susceptibility genes from GWAS, whose roles in disease susceptibility are not previously understood, and applied rcNet algorithms to predict the disease phenotype of the disease susceptibility genes. We extracted all the disease susceptibility genes discovered in GWAS based on a recent survey of all studies reported in the GWAS catalog as of Dec. 2010 [22]. After filtering out the genes already included in OMIM May-2007 Version, we selected 217 diseases/traits with novel susceptibility genes that are not associated with any disease phenotype in OMIM May-2007 Version, and 31 out of the 217 diseases/traits could be matched with OMIM phenotypes in the disease network. Subsequently, the 31 diseases/traits and their susceptibility genes were used in this experiment.

We queried the set of disease susceptibility genes of each of the 31 diseases/traits to rank the 5080 OMIM disease phenotypes. A disease/trait could be matched with multiple OMIM disease phenotypes. We report the rank of the matched phenotype with the best ranking for each query. The ranking results of a subset of the 31 diseases/traits are reported in table 2. Among the 31 queries, 14 cases ranked the target diseases within top 2% (ranked within top 100). Notable examples are prostate cancer, breast cancer, basal cell carcinoma, bladder cancer, acute lymphoblastic leukemia, systemic lupus erythematosus, and leprosy. In these cases, the rcNet algorithms ranked the target disease phenotype of the query gene set within top 1%. Fig. 6 shows the example that rcNet accurately ranked the breast cancer phenotypes of the breast cancer susceptibility genes, by querying with 26 novel breast cancer susceptibility genes from GWAS. The connectivity between the top ranked disease genes and the top ranked disease phenotypes is around 13 folds of the expected number of connections between the same numbers of

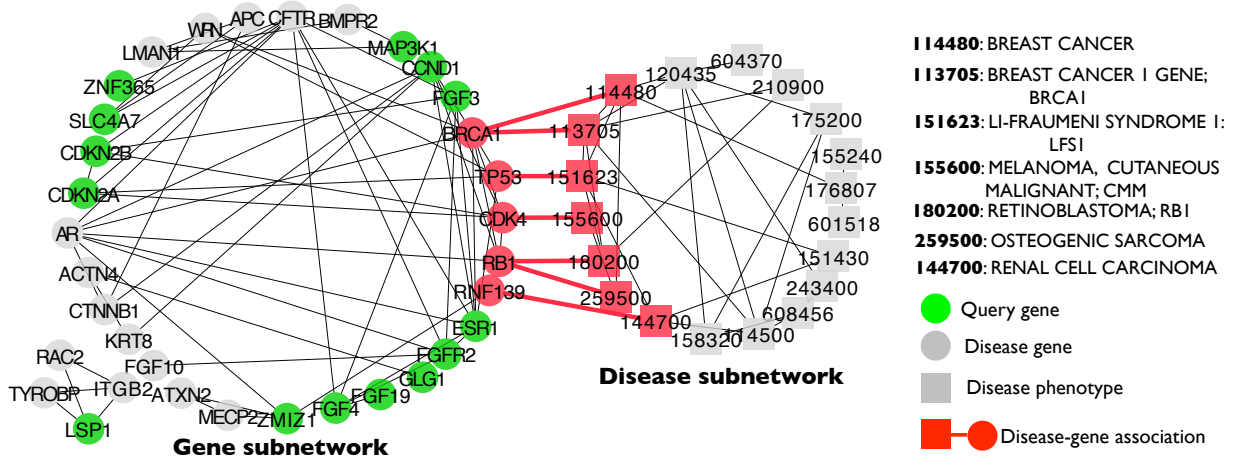


Figure 6. Querying with breast cancer susceptibility genes from GWAS by rcNet. This example shows how rcNet algorithm predicted the target disease phenotypes of breast cancer susceptibility genes from GWAS. By querying with the 26 novel breast cancer susceptibility genes from GWAS, rcNet ranked the 20 disease genes in the gene subnetwork at the top. The gene subnetwork also includes 14 out of the 26 query genes, which are connected with the top-20 genes. Similarly, the top 20 disease phenotypes ranked by OMIM114480:breast cancer disease phenotype are included in the disease subnetwork. In this example, 5 of the 20 top-ranked disease genes are connected to 7 of the top-20 disease phenotypes given by 7 OMIM disease-gene associations, compared with the expected 0.87 association between 34 random genes and 20 random phenotypes.

random genes and phenotypes, or around 11 folds of the average number of connections given from the phenotype ranking by the relevance to the unrelated disease phenotypes. One interesting observation is that the target disease phenotype OMIM:113705 ‘BREAST CANCER 1 GENE; BRCA1’ is only directly connected with two top ranked disease phenotypes, OMIM: 114480 ‘BREAST CANCER’, OMIM:151623 ‘LI-FRAUMENI SYNDROME 1:LFS1’, and OMIM:‘259500: OSTEOGENIC SACROMA’, and only 5 of the 26 query genes directly interact with the top ranked disease genes with disease-gene associations. The neighbor expansions in both the gene network and the phenotype network resulted in 4 OMIM disease-gene associations. This observation suggests that, simply exploring the direct neighbors of the query gene set and the target disease phenotype in the networks, a method might fail to infer disease-gene set associations, due to the low statistical significance of the sparse connectivity between the genes and the disease phenotypes. Specifically, in this example, the fold enrichment for 4 associations is 7.53, which is significantly lower than the 12.35 fold enrichment obtained by rcNet. Another interesting example is the inference of the association between leprosy and its susceptibility genes from GWAS (pubmed 20018961). In OMIM May-2007 Version, leprosy has no causative genes, and the leprosy susceptibility genes from GWAS also have no association with any disease. The lack of known associations in both the target disease phenotype and the get set poses a hard case that gene set enrichment analysis based on overrepresentation will fail to reveal, but rcNet algorithms ranked Leprosy within top 2%.

In contrast to the results in cross-validation on OMIM data, rcNet produced significantly better results in cancer, immunological, and gastrointestinal disease, compared with rcNet_{corr} and rcNet_{lap}. Interestingly, previous studies showed that disease susceptibility genes from GWAS catalog have less modularities in the gene network compared to those of the known disease genes in OMIM, and phenotypically similar diseases such as immunological and gastrointestinal diseases do not tend to share their disease genes [23, 24]. Those previous studies also implicated that due to the unique topological characteristics of disease susceptibility genes discovered in GWAS, the existing network-based methods would

fail to reveal the associations between the disease susceptibility genes and the disease. However, our experiments suggest that, by incorporating the global topological information in the networks and the known OMIM associations, rcNet algorithms can successfully discover the elusive associations in many cases.

Predicting Disease Phenotypes of Genes with Copy Number Changes

In DNA copy number analysis, genes in the chromosomal regions with copy number changes are identified as candidate disease genes. In this experiment, we applied the rcNet algorithms to predict the target disease phenotypes of the candidate disease genes in disease susceptible copy number change regions. We collected 13 human DNA copy number change datasets from a recent human cancer copy number study from <http://www.broadinstitute.org/tumorscape> [25]. The DNA copy number measurements in the datasets were obtained on the Affymetrix 250K Sty SNP array. The regions with copy number changes were detected by GISTIC tool with default settings [26]. Genes in the detected copy number change regions were used as the query gene set to predict their target disease phenotypes.

Table 3 shows the ranking results by the rcNet algorithms. rcNet ranked the target disease within top 2% for 6 of the 13 cancers and rcNet_{corr} ranked the target disease within top 2% for 7 of the 13 cancers. In 9 of the cases, at least one algorithm ranked the target disease within top 100. [25] stated that more than three-quarters of the statistically significantly altered copy number regions contain potential cancer causing genes that are not previously validated targets of somatic copy number alternations in human cancer. This suggests that enrichment analysis of the genes will not reveal any disease-association, but rcNet algorithms found many associations with the network information.

Table 3. Ranking the target disease phenotypes of the candidate disease genes with copy number changes. This experiment includes 13 human cancer copy number studies from [25].

Disease/Trait	Rank by rcNet	Rank by rcNet _{corr}	Rank by rcNet _{jap}
Neuroblastoma	5	13	126
Colorectal cancer	14	20	613
Renal cancer	22	14	33
Non small cell lung cancer	34	48	558
Breast cancer	68	136	521
Medulloblastoma	77	826	2007
Prostate cancer	129	127	2447
Ovarian cancer	322	73	1108
Small cell lung cancer	759	53	909
Mesothelioma	959	21	54
Gastrointestinal stromal tumor	1169	787	1679
Hepatocellular carcinoma	4241	952	1295
Glioma	4705	787	951

Predicting Disease Phenotypes of Differentially Expressed Genes

It is frequently observed that many disease susceptibility genes are not differentially expressed in microarray gene expression experiments. In this experiment, we applied rcNet algorithms to predict the target disease of differentially expressed genes in gene expression profiles. We collected 13 human cancer microarray gene expression dataset from GEO. Gene expression profiles were obtained on the Affymetrix HU133A array, and normalized by RMA [27]. Standard *t*-test was used to identify differentially expressed genes. The differentially expressed genes were used to query for their target diseases. To quantify how reliable a differential expression is, the query gene nodes were initialized by the absolute values of the *t*-statistics for label propagation. Table 4 reports the results of predicting the target diseases of the differentially expressed genes. Out of the 13 cases, the rcNet algorithms could rank 7 within top 5%, and 12 within top 10%. Although the result is only moderately encouraging, it validates the hypothesis that the

neighboring information of the differentially expressed genes provides clue of association with the target disease phenotype.

Table 4. Ranking the target disease of differentially expressed genes. The first column represents the target disease of a microarray gene expression study, and the second column gives the GEO number of the dataset.

Disease/Trait	GEO Num.	Rank by rcNet	Rank by rcNet _{corr}	Rank by rcNet _{lap}
AML	GSE9476	576	316	359
Breast cancer	GSE7390	14	49	51
	GSE2034	40	130	146
	GSE6532	129	151	182
	GSE1456	138	102	109
	GSE3494	161	709	1313
Gastric cancer	GSE13911	248	298	362
Lung cancer	GSE10072	206	755	2219
	E-MEXP-231	318	608	1115
	GSE7670	379	1330	4002
Ovarian cancer	GSE6008	414	1494	2283
Prostate cancer	E-MEXP-1327	271	1446	2057
	GSE8218	900	1214	2498


rcNet WebTool

The rcNet algorithms were implemented and deployed as a general webtool for disease-gene set association analysis at http://compbio.cs.umn.edu/dgsa_rcNet. Providing a list of query genes, a user can retrieve the OMIM disease phenotypes ranked by their degree of association with the gene set. In Fig. 7, we show an example of querying rcNet WebTool with the disease gene set of prostate cancer from GWAS. In the implementation, the Laplacian scores are precomputed to improve efficiency. Currently, it takes the server less than 5 seconds to response to a gene set query.

Computational Biology Lab

Department of Computer Science and Engineering, University of Minnesota - Twin Cities

Introduction | User Guidance | Reference | Credits



rcNet: A Web Tool for Inferring Disease and Gene Set Association

Gene Set Query | **Phenotype Query**

Enter your genes into the text area:
(One gene in one line)

C2orf43
CTBP2
EHBP1
FOXP4
GPRC6A
GSPT2
JAZF1
KLK3
LMTK2
MAGED1

Search gene:

Select a ranking method:
rcNet_lap(Laplacian Score)

The number of phenotypes to display:
10

View query gene set:

TNRC6B | SLC22A3 | LMTK2 | MAGED1 | JAZF1 | GPRC6A | FOXP4 | KLK3 | MSMB | CTBP2 | EHBP1 | NUDT10 | NUDT11 | GSPT2 | LESS

Phenotype ranking:

Ranking	Phenotype Name	Relevance Score
1	TETRALOGY OF FALLOT	1273.4128
2	PROSTATE CANCER	1439.6707
3	IMMUNODYSREGULATION, POLYENDOCRINOPATHY, AND ENTEROPATHY, X-LINKED, IPEX	1494.3404
4	PROGEROID FACIAL APPEARANCE WITH HAND ANOMALIES	1556.6287
5	MESOAXIAL HEXADACTYLY AND CARDIAC MALFORMATION	1614.4134
6	ECTRODACTYLY OF LOWER LIMBS, CONGENITAL HEART DEFECT, AND MICROGNATHIA	1629.6243
7	SPINOCEREBELLAR ATAXIA 2, SCA2	1640.2059
8	MICROCEPHALY-CARDIOMYOPATHY	1652.7193
9	OVARIAN GERM CELL CANCER	1654.8965
10	BRACHYMORPHISM-ONYCHODYSPLASIA-DYSPLALANGISM SYNDROME	1688.1062

[About Us](#) | [Site Map](#) | [Privacy Policy](#) | [Contact Us](#) | ©2011 Computational Biology Lab, UMN

Figure 7. rcNet WebTool Demo. In this example, a gene set with a list of 15 genes identified as prostate cancer susceptibility genes in GWAS was used to query rcNet WebTool. The left panel shows the settings used for query and the right panel displays the query result.

Discussion

Analysis of the gene sets from genome-wide high-throughput screening is a continuing challenge in many disease studies. When the gene set is poorly annotated, enrichment analysis will fail to detect any associations with disease phenotypes, or when the gene set contains genes in a broad range of functional categories, enrichment analysis provides unreliable statistical significance. Statistics from OMIM (Jan 2011) show that 3745 of the 6675 disease phenotypes are still unknown for their molecular basis. Thus, enrichment analysis will fail to find any associations between the 3745 disease phenotypes and any query gene set. For example, in the experiments with the GWAS gene sets, rcNets algorithms ranked leprosy (OMIM:246300 and OMIM:607572) among the top 2% phenotypes, while enrichment analysis reported no association for the four disease susceptibility genes of leprosy. rcNet focuses on improving detection of disease phenotype-gene set associations by integrating gene network and disease network to better summarize sparse associations for a global comparison of all possible disease and gene set associations. The rcNet algorithms effectively utilizes hidden information in the gene network and the disease network with the machine learning models. First, the label propagation steps on both the gene network and the disease network fully explore the neighborhood information of the query genes and a disease phenotype. The relevance information is propagated from the seed nodes to their neighbors to provide a global quantification of relevance, and the relevance scores are then utilized with all the known associations for evaluating the association between the gene set and the disease phenotype. Thus, analysis with rcNet is not biased by poor known annotation or the size of the query gene set. Second, compared with the other methods that also utilizing the gene network and the disease network, rcNet is more flexible in handling the network data because rcNet is capable of handling weighted associations and weighted edges in the gene network and the disease network. rcNet does not rely on deciding direct neighbors or shortest path as CIHPER or PRINCE [15]. Finally, the ridge regression model coupled with label propagation provides an approximation of finding association between a gene set and multiple disease phenotypes, which is difficult to achieve with enumeration-based strategies.

Despite the encouraging results of rcNet, there are also limitations. First, rcNet relies heavily on the networks. For the cases where the gene set already has known associations with the target disease phenotype, the network information might introduce noise to dilute the strong signal as showed in Fig. 5. Thus, rcNet is more useful for studying new diseases that have not been genetically characterized rather than confirming well-understood diseases. Second, it is also difficult to distinguish the closely related phenotypes from false positives, because it is possible that some of the top-ranked phenotypes is not similar to or share any common disease genes with the target disease phenotype in the disease network. Interpretation of these phenotypes will not be straightforward. A possible solution is to identify subnetworks as the example in Fig. 6 and use information from the gene cluster and the phenotype cluster for finding explanations.

rcNet is a helpful tool for oncologists to analyze disease and gene set association. Researchers can validate their findings from high-throughput studies, especially, to validate novel associations between complex diseases and a query gene set with no known associations. rcNet can also help identify closely related phenotypes of the target disease of the query gene set. Since rcNet algorithms utilize both the disease similarities and the gene interactions, some phenotypically similar disease phenotypes will be ranked at the top. These disease phenotypes might provide additional information to investigate the target disease in the study. In future work, we plan to extend rcNet as a more general tool that could also infer enriched functions of a query gene set. We can build a GO function network with GO term nodes and edges weighted by the similarity between two GO functions. The application of the rcNet algorithm under this context will be straightforward.

Acknowledgments

This work is partially supported by IHI Research Seed Grant from the Institute of Health Informatics at University of Minnesota TC.

References

1. McKusick V (2007) Mendelian inheritance in man and its online version, omim. *Am J Hum Genet* 80: 588–604.
2. The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447: 661-678.
3. Johnson A, O'Donnell C (2009) An open access database of genome-wide association results. *BMC Med Genet* 10: 6.
4. Shlien A, Malkin D (2009) Copy number variations and cancer. *Genome Med* 1: 62.
5. van't Veer L, Bernards R (2008) Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature* 452: 564-570.
6. Huang D, Sherman B, Lempicki R (2009) Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat Protoc* 4: 44-57.
7. Subramanian A, Tamayo P, Mootha V, Mukherjee S, Ebert B, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102: 15545-15550.
8. Martin D, Brun C, Remy E, Mouren P, Thieffry D, et al. (2004) GOTOolbox: functional analysis of gene datasets based on gene ontology. *Genome Biol* 5: R101.
9. van Driel M, et al. (2006) A text-mining analysis of the human phenome. *Eur J Hum Genet* 14: 535–542.
10. Wu X, et al. (2008) Network-based global inference of human disease genes. *Mol Syst Biol* 4.
11. Linghu B, Snitkin E, Hu Z, Xia Y, Delisi C (2009) Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome Biol* 10: R91.
12. Franke L, van Bakel H, Fokkens L, de Jong E, Egmont-Petersen M, et al. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet* 78: 1011-1025.
13. Khler S, et al. (2008) Walking the interactome for prioritization of candidate disease genes. *The American J of Hum Genet* 82: 949 - 958.
14. Hwang T, Kuang R (2010) A heterogeneous label propagation algorithm for disease gene discovery. *Proc of SIAM International Conference on Data Mining* : 583-594.
15. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R (2010) Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol* 6: e1000641.
16. Li Y, Patra J (2010) Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics* 26: 1219-1224.

17. Bengio Y, et al. (2006) Label propagation and quadratic criterion. In: Chapelle EO, et al., editors, *Semi-Supervised Learning*, MIT Press.
18. Zhou D, et al. (2004) Learning with local and global consistency. In: *NIPS*. volume 16, pp. 321-328.
19. Peri S, Navarro J, Amanchy R, Kristiansen T, Jonnalagadda C, et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* 13: 2363-2371.
20. Huttenhower C, Haley EM, Hibbs MA, Dumeaux V, Barrett DR, et al. (2009) Exploring the human genome with functional maps. *Genome Research* 19: 1093–1106.
21. Goh K, et al. (2007) The human disease network. *Proc Natl Acad Sci USA* 104: 8685–8690.
22. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 106: 9362.
23. Baranzini SE (2009) The genetics of autoimmune diseases: a networked perspective. *Current opinion in immunology* 21: 596–605.
24. Barrenas F, Chavali S, Holme P, Mobini R, Benson M (2009) Network properties of complex human disease genes identified through genome-wide association studies. *PLoS ONE* 4: e8090.
25. Beroukhi R, Mermel CH, Porter D, Wei G, Raychaudhuri S, et al. (2010) The landscape of somatic copy-number alteration across human cancers. *Nature* 463: 899–905.
26. Beroukhi R, Getz G, Nghiemphu L, Barretina J, Hsueh T, et al. (2007) Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proc Natl Acad Sci* 104: 20007–20012.
27. Irizarry R, Hobbs B, Collin F, Beazer-Barclay Y, Antonellis K, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4: 249.